

Multi-View Facial Expression Recognition with AU-Guided Heatmaps

Hong-Quan Do^{1,2}, Hoang Van Thanh³, and Tu Minh Phuong^{2,*}

¹FPT University, Hanoi, Vietnam

²Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

³Quang Binh University, Quang Tri, Vietnam

Email: quandh13@fe.edu.vn, thanhhv@qbu.edu.vn, phuongtm@ptit.edu.vn

Abstract—Facial Expression Recognition (FER) is fundamental to affective computing and human-computer interaction, enabling systems to infer emotional states from facial cues. Current deep learning approaches rely solely on raw facial images and standard augmentation but lack explicit incorporation of facial anatomical knowledge. This forces models to discover expression-discriminative regions from scratch rather than leveraging established facial behavior research. To address this gap, we propose MV-FER, a multi-view framework that integrates spatial priors from the Facial Action Coding System (FACS). MV-FER employs three data representations: original facial images, horizontally flipped images with random erasing for view consistency, and AU-guided heatmaps generated by mapping landmarks to established Action Unit areas providing explicit spatial guidance toward expression-relevant facial regions. The three views are unified within a training paradigm that encourages learning of discriminative features while maintaining robustness against pose variations and partial occlusions. Benchmark experiments demonstrate that MV-FER achieves enhanced recognition accuracy over state-of-the-art approaches with similar architectures. These results confirm the effectiveness of our multi-view strategy, with progressive accuracy gains as augmented and AU-guided heatmap views are incorporated.

Index Terms—Action Unit, Facial Action Coding System, Facial Expression Recognition, Facial Landmark, Heatmap, Multi View.

I. INTRODUCTION

The ability to recognize and interpret human facial expressions represents one of the most fundamental aspects of human social interaction and emotional communication. As computational systems become increasingly integrated into human-centered applications, the automatic recognition of facial expressions has emerged as a critical enabling technology for affective computing, human-computer interaction, and intelligent user interfaces. Facial Expression Recognition (FER) systems hold immense potential across diverse domains, including mental health monitoring, educational technology, customer experience analysis, and human-robot interaction, where understanding human emotional states can significantly enhance system responsiveness and user experience. The central challenge in FER lies in learning robust and discriminative feature representations despite significant intra-class variations

induced by pose changes, illumination conditions, partial occlusions, and subtle intensity differences between expressions.

The advancement of deep learning techniques has revolutionized the field of computer vision, and FER has benefited substantially from these developments. Deep learning approaches for facial expression recognition have demonstrated remarkable progress, primarily leveraging convolutional neural networks (CNNs) and their variants to learn hierarchical feature representations directly from raw facial images [1], [2]. These methods typically employ standard data augmentation strategies including geometric transformations, color space modifications, and noise injection to improve model generalization and enhance robustness against variations in imaging conditions. Although these computational pipelines have achieved considerable progress, their reliance on single-view representations often constrains their ability to generalize under challenging real-world conditions.

To address these limitations, recent studies have increasingly adopted multi-view learning paradigms. In such approaches, multiple augmented versions of the same input image—obtained through transformations including horizontal flipping, random cropping, affine warping, or color space modifications—are treated as distinct yet semantically consistent views. Joint learning across these complementary views enhances model robustness and encourages invariance to nuisance factors. However, conventional multi-view augmentation strategies remain generic in nature, as they can be applied to any image recognition task without exploiting the unique anatomical structure of human faces.

In contrast, FER presents a distinctive opportunity to leverage domain-specific priors derived from established psychological and anatomical frameworks. Specifically, Action Units (AUs), as formally defined by the Facial Action Coding System (FACS) developed by Ekman and Friesen [3], describe localized regions of facial muscle activity that exhibit strong correlations with emotional expressions. Prior research [4], [5] has established comprehensive mappings between facial expressions and their corresponding action units. For instance, the expression 'Surprise' is systematically characterized by AU1 (inner brow raiser), AU2 (outer brow raiser), AU5 (upper lid raiser), and AU26 (jaw drop), with analogous detailed mappings documented for other basic emotions. By integrating

*Corresponding author.

an AU codebook with facial landmark detection and expression labels, it becomes feasible to generate AU-guided heatmaps that systematically highlight the most discriminative facial regions—including the brows, eyes, and mouth areas—relevant to specific expressions. These heatmaps constitute semantically meaningful, structured representations that emphasize expression-critical anatomical areas. We propose that AU-guided heatmaps represent an additional, FER-specific view of the input data, complementing conventional augmented images while providing enhanced guidance for discriminative feature learning.

In this paper, we develop a multi-view facial expression recognition framework that integrates domain-specific knowledge from FACS theory into deep learning architectures. Our approach utilizes three complementary views: original facial images and horizontally flipped augmented images as dual input branches, while AU-guided heatmaps—derived from facial landmarks and FACS theory—to guide attention toward expression-critical regions including the eyes, brows, and mouth areas. The framework employs a unified training paradigm ensuring consistent representation learning across both input views while aligning learned attention with expression-relevant facial regions. This integration of multi-view learning and attention alignment enables our proposed method, named MV-FER, to achieve enhanced facial expression recognition performance.

The key contributions of this work can be summarized as follows:

- **AU-guided heatmap Generation:** We introduce a method for generating AU-guided heatmaps from facial landmarks and FACS theory, which serve as a signal to guide the spatial attention mechanisms of deep networks toward expression-based regions.
- **Multi-view learning architecture:** We propose MV-FER, a unified training paradigm that synthesizes three data views: original images, augmented flipped images with random erasing for view consistency, and AU-guided heatmaps for attention alignment.
- **Experimental validation:** We demonstrate consistent performance improvements over state-of-the-art methods on benchmark dataset RAF-DB, with ablation studies confirming the progressive gains from each view component.

II. RELATED WORKS

Recent advances in Facial Expression Recognition (FER) have increasingly incorporated domain-specific knowledge to enhance model performance. This section reviews key developments in facial landmark utilization, action unit-guided supervisions, and multi-view learning approaches that form the foundation for our work.

A. Facial Landmarks for Expression Recognition

Facial landmark detection, which identifies and localizes predefined keypoints on human faces, has become instrumental in numerous face analysis applications. In FER, researchers have leveraged facial landmarks as informative structural

features to complement raw image data. Jung et al. [6] pioneered this direction by introducing a dual-network CNN architecture that processes both facial images and landmarks through weighted summation. Building on this foundation, [7] proposed a dual-branch network combining CNN and Graph Convolutional Networks (GCN), achieving notable results on benchmark datasets including CK+ (99.24%), FER2013 (73.26%), and RAF-DB (87.42%). More recent transformer-based approaches have further advanced landmark utilization in FER, including POSTER [8] and POSTER++ [9]. Notably, both POSTER variants extract landmark-inspired features from pre-trained models rather than using explicit landmark coordinates, demonstrating the value of landmark-derived representations even when landmarks are not directly accessible.

B. Action Unit-Guided Supervision

Beyond direct landmark utilization, recent research has explored heatmap-based approaches to guide model attention toward expression-relevant facial regions. However, instead of creating heatmap from all landmarks, a particularly promising direction involves leveraging Action Units (AUs) from the Facial Action Coding System (FACS) to create targeted supervision signals. AUs represent specific facial muscle movements associated with emotional expressions, providing expert-defined anatomical knowledge that can guide deep learning models. [3], [4] researches that each basic expression corresponds to specific AU combinations localized to particular facial regions. Recently, in 2024 [5] explicitly integrated spatial AU cues into the training process by constraining layer-wise attention features of FER classifiers to align with AU heatmaps. This approach ensures that models focus on expert-defined regions of interest, demonstrating improved interpretability and performance on RAF-DB and AffectNet datasets. However, this work primarily focuses on single-view supervision without exploring the potential of multi-view learning paradigms.

C. Multi-view Learning Approaches

Multi-view learning represents an emerging paradigm that enhances learning performance by leveraging multiple views or sources of data. By integrating information from diverse perspectives, multi-view learning methods effectively improve accuracy, robustness, and generalization capabilities. The foundational approach of multi-view learning can be traced to co-training algorithms, where two initial predictive functions f_1 and f_2 are trained on separate views using labeled data. The algorithm proceeds iteratively: the most confident examples predicted by f_1 are incorporated into the labeled set to retrain f_2 , and vice versa, with both functions subsequently retrained on the expanded labeled datasets [10]. While general multi-view learning has witnessed extensive development across various domains, its application to facial expression recognition presents distinctive opportunities and challenges that warrant specialized approaches.

In the context of multi-view facial expression recognition, several pioneering methods have demonstrated the effective-

ness of leveraging multiple facial regions or perspectives. The pACNN approach [11] extracts 24 regions of interest from intermediate feature maps, employing a Patch-Gated Unit (PG-Unit) for each patch region to weight local representations according to the region's occlusion status—specifically, the extent to which each patch remains unobstructed. These weighted representations are subsequently concatenated and fed to the classification component. Building upon this local analysis, gACNN integrates weighted local representations with global facial representations, where the global features are encoded and weighted through a Global-Gated Unit (GG-Unit). Similarly, the Region Attention Network (RAN) [12] adopts a complementary strategy by cropping facial images into multiple regions using three distinct approaches: fixed position cropping, random cropping, and landmark-based cropping. These extracted regions are processed through a backbone CNN for feature extraction, followed by self-attention and relation-attention modules to derive compact facial representations. This multi-region approach demonstrates the value of considering different facial areas as distinct but complementary views. Recent research has further explored the concept of view consistency in facial expression analysis. Studies have revealed that flip semantic consistency of facial expression images can serve as an effective mechanism for detecting noisy labels [13]. The principle of flip semantic consistency posits that an original facial image and its horizontally flipped counterpart should be classified into the same emotional category, providing a natural form of data augmentation and consistency constraint for multi-view learning frameworks.

Building on these insights, our work introduces a novel multi-view learning framework that differs in several key ways. First, instead of using landmark-derived heatmaps as direct input features, we generate AU-guided heatmaps that serve purely as supervisory signals to guide attention learning—essentially teaching the network where to focus on expression-critical regions. Second, we employ a dual-branch architecture where each facial image is processed alongside its horizontally flipped counterpart as complementary views. This design leverages flip-semantic consistency—the principle that original and flipped facial expressions should yield identical classifications—an approach that remains largely underexplored in multi-view FER. Together, this creates a three-view learning paradigm that enriches data diversity while embedding domain-specific priors.

III. PROPOSED METHOD

A. Framework Overview

We propose MV-FER, a multi-view framework that integrates original images, augmented views, and AU-guided heatmaps to improve FER performance as depicted in Figure 1.

1) *Training Phase*: The training paradigm jointly optimizes expression classification, cross-view consistency, and spatial attention alignment with expert-defined facial regions. Given an input image X , we construct three views:

- Original view: $x_1 = X$

- Augmented view: $x_2 = \mathcal{T}(X)$, where $\mathcal{T}(\cdot)$ applies horizontal flipping with random erasing to encourage view-consistent learning.
- AU-guided heatmap: $x_3 = \mathcal{H}(X) \in [0, 1]^{H \times W}$, where $\mathcal{H}(\cdot)$ generates spatial emphasis maps highlighting expression-relevant facial regions based on FACS.

During training, the pair (x_1, x_2) is fed into a shared backbone network $\mathcal{F}(\cdot; \theta)$ and Classifier head $\mathcal{G}(\cdot; \phi)$ to produce class predictions and feature representations, while x_3 is used for attention alignment. In the validation and testing phases, only the original images are utilized. Conversely, in the validation and testing phases, only the original view x_1 are employed.

The overall training objective is defined as a weighted combination of three losses:

- **Classification Loss** (cross-entropy for original branch):

$$\mathcal{L}_{\text{cls}} = - \sum_{c=1}^C y_c \log \hat{p}_{1,c} = \text{CE}(\hat{\mathbf{p}}_1, \mathbf{y}) \quad (1)$$

where y_c is the one-hot ground-truth label for class c , and $\hat{\mathbf{p}}_1 = \text{Softmax}(\mathcal{G}(\mathcal{F}(x_1)))$

- **Consistency Loss** (between CAMs of original branch and augmented branch): First we extract class activation maps (CAM) [14] after the shared backbone network $\mathcal{F}$:

$$A = \text{CAM}(\mathcal{F}(x_1)), \quad A' = \text{CAM}(\mathcal{F}(x_2)) \quad (2)$$

Second since x_2 is undergo geometric transformations, A' is spatially mapped back into the grid of A :

$$\tilde{A}' = \mathcal{W}_{\tau^{-1}}(A') \quad (3)$$

Then minimize the distance between A and $\tilde{A}'$

$$\mathcal{L}_{\text{con}} = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W D(A_{ij} \| \tilde{A}'_{ij}) \quad (4)$$

where $D(\cdot \| \cdot)$ denotes the squared error by default, but can be replaced with other divergence or distance measures.

- **Alignment Loss**: Let $\bar{A} = \text{norm}(A)$ and $\bar{H} = \text{norm}(x_3)$ denote spatial normalizations to $[0, 1]$ (and optionally $\sum \bar{A} = \sum \bar{H} = 1$ for probabilistic divergences). Then:

$$\mathcal{L}_{\text{align}} = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W D(\bar{A}_{ij} \| \bar{H}_{ij}) \quad (5)$$

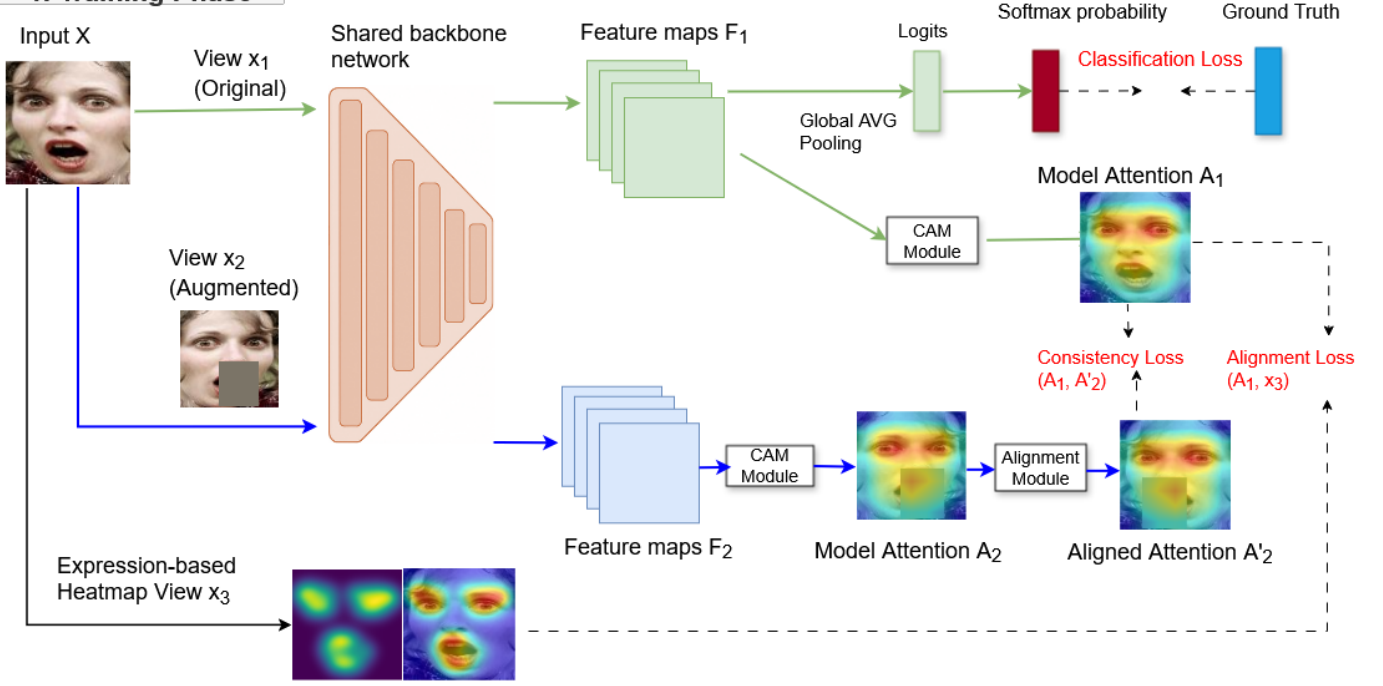
The final training objective is a weighted combination of the three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda * \mathcal{L}_{\text{con}} + \beta * \mathcal{L}_{\text{align}} \quad (6)$$

where λ and β are balancing hyperparameters.

2) *Inference Phase*: During the inference phase, only the original view of each image in the test set is employed. Each image is fed into the trained backbone network to generate class logits, which are subsequently passed through a softmax layer to obtain a probability distribution over the seven expression categories: surprise, fear, disgust, happiness, sadness, anger, and neutral. The final prediction is determined by selecting the class with the highest probability.

1. Training Phase



2. Inference Phase

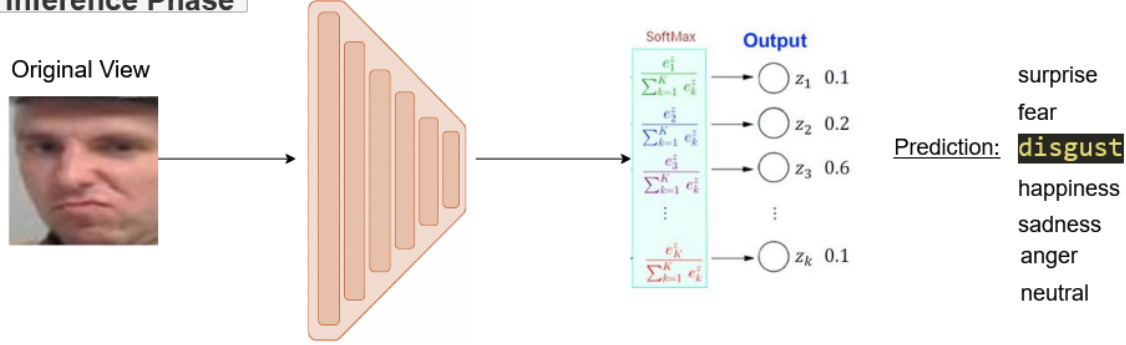


Fig. 1: Overall architecture of the proposed method MV-FER.

B. AU-Guided Heatmap Generation

The Facial Action Coding System (FACS) [3] provides an anatomically grounded framework for decomposing facial expressions into Action Units (AUs)—elementary muscle movements that form the basis of emotional displays. Each basic emotion is characterized by a distinctive subset of AUs localized to specific facial regions, enabling systematic computational analysis of expressions.

This established knowledge enables us to construct AU-guided heatmaps that direct attention toward expression-relevant areas. Prior research has thoroughly documented the mappings between emotions and their constituent AUs [3]–[5], which we summarize in Tables I and II. For clarity, note that AU26 (Jaw Drop) is anatomically defined as the relaxation of the masseter, temporalis, and medial pterygoid muscles that lower the mandible, primarily affecting mouth opening rather than the visible jawline contour. By mapping

facial landmarks to these documented AU areas, we generate spatial guidance that directs model attention toward the most informative regions for each expression.

TABLE I: Basic Facial Expressions with Corresponding Action Units based on prior research [3]–[5]

Expression	AU	FACS Definition
Happiness	6, 12, 25	Cheek Raiser, Lip Corner Puller, Lips Part
Surprise	1, 2, 5 26, 27	Inner Brow Raiser, Outer Brow Raiser Upper Lid Raiser, Jaw Drop, Mouth Stretch
Fear	1, 2, 4 5, 20, 25 26, 27	Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer Upper Lid Raiser, Lip Stretcher, Lips Part Jaw Drop, Mouth Stretch
Anger	4, 5, 7 10, 17, 23 24, 25, 26	Brow Lowerer, Upper Lid Raiser, Lid Tightener Upper Lip Raiser, Chin Raiser, Lip Tightener Lip Presser, Lips Part, Jaw Drop
Disgust	9, 10, 16 17, 25, 26	Nose Wrinkler, Upper Lip Raiser, Lower Lip Depressor Chin Raiser, Lips Part, Jaw Drop
Sadness	1, 4, 6 15, 17	Inner Brow Raiser, Brow Lowerer, Cheek Raiser Lip Corner Depressor, Chin Raiser

Figure 2 illustrates our pipeline of generating AU-Guided

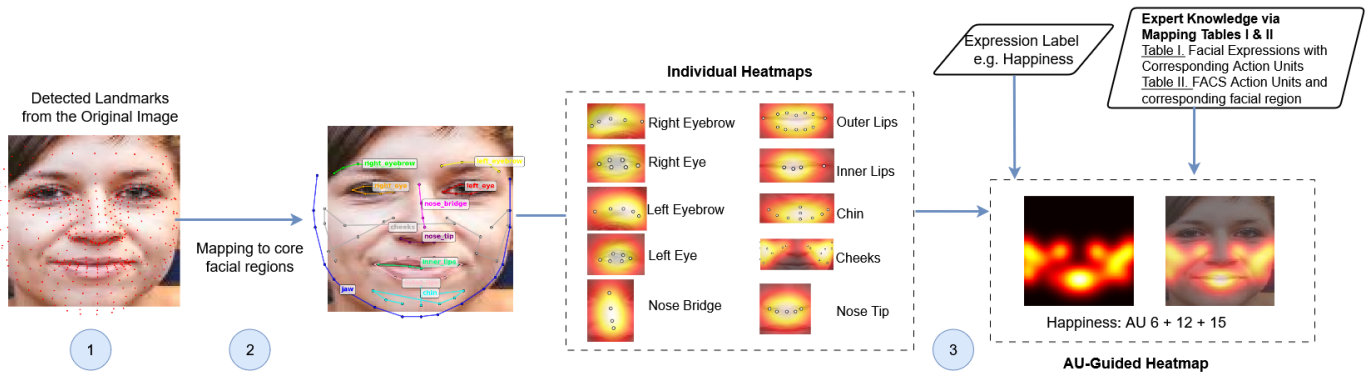


Fig. 2: AU-Guided Heatmap Generation Pipeline

TABLE II: FACS Action Units and corresponding facial regions.

AU	FACS Definition	Primary Facial Region
1	Inner brow raiser	Left eyebrow, Right eyebrow
2	Outer brow raiser	Left eyebrow, Right eyebrow
4	Brow lowerer	Left eyebrow, Right eyebrow, Nose bridge
5	Upper lid raiser	Left eye, Right eye
6	Cheek raiser	Cheeks, Left eye, Right eye
7	Lid tightener	Left eye, Right eye
8	Lips toward each other	Inner lips, Outer lips
9	Nose wrinkler	Nose bridge, Nose tip, Cheeks
10	Upper lip raiser	Outer lips, Inner lips, Cheeks
11	Nasolabial deepener	Cheeks
12	Lip corner puller	Outer lips, Cheeks
13	Sharp lip puller (caninus)	Outer lips, Cheeks
14	Dimpler	Cheeks, Outer lips
15	Lip corner depressor	Outer lips
16	Lower lip depressor	Outer lips, Inner lips, Chin
17	Chin raiser	Chin, Inner lips
18	Lip pucker	Inner lips, Outer lips
19	Tongue show	Inner lips
20	Lip stretcher	Outer lips, Cheeks
21	Neck tightener	Jaw
22	Lip funneler	Inner lips, Outer lips
23	Lip tightener	Outer lips
24	Lip pressor	Inner lips, Outer lips
25	Lips part	Inner lips
26	Jaw drop	Inner lips, Outer lips
27	Mouth stretch	Inner lips, Outer lips
28	Lip suck	Inner lips, Outer lips

heatmaps. The process initiates with facial landmark detection on the input image, providing precise spatial coordinates for key facial features. These detected landmarks are subsequently mapped to anatomically meaningful facial regions that correspond to expression-critical areas as defined by FACS theory. The systematic construction of expression-specific heatmaps is enabled through the integration of expert knowledge via pre-established lookup tables (Tables I and II). These tables provide comprehensive mappings between facial expressions and their corresponding Action Units, as well as the spatial relationships between AUs and specific facial regions.

The framework generates expression-based heatmaps by combining relevant regional activations according to the AU composition of target expressions. For instance, the happiness expression is systematically constructed by activating regions associated with AU6 (cheek raiser) in the cheek and eye areas, AU12 (lip corner puller) and AU15 (lip corner depressor)

in the cheek and outer lip regions. This process results in composite expression-specific heatmaps.

IV. EXPERIMENTAL SETUP

A. Dataset

We evaluate the effectiveness of our proposed method on the Real-world Affective Faces Database (RAF-DB) [15], a widely adopted benchmark for facial expression recognition. RAF-DB is a large-scale, crowd-annotated dataset that consists of 29,672 real-world RGB facial images, each with a resolution of (100×100) pixels. The dataset has been carefully curated to cover diverse conditions, including variations in age, gender, ethnicity, head pose, illumination, and occlusion, making it highly representative of real-world FER challenges.

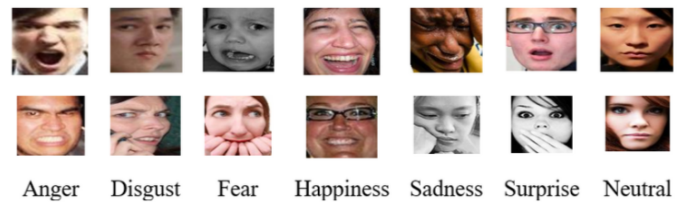


Fig. 3: Sample images in RAF-DB [15] dataset.

Each image in RAF-DB is annotated either with one of seven basic emotion categories (happiness, surprise, fear, anger, disgust, sadness, and neutral) or, in the case of the two-label subset, with one of twelve compound emotion categories that capture more nuanced affective states. In this study, we restrict our evaluation to the single-label subset, as it represents the standard benchmark setting in facial expression recognition (FER) research. Following the official partitioning protocol, we use 12,271 images for training—of which 20% are held out for validation—and 3,068 images for testing. It is important to note that the dataset is highly imbalanced, with an imbalance ratio of 16.78, reflecting substantial variation in class frequencies as shown in 4.

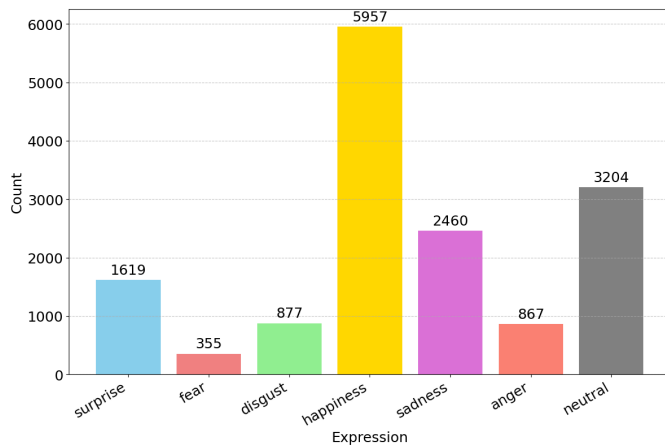


Fig. 4: RAF-DB Class Distribution.

B. Implementation Detail

In our experimental evaluation, we employed ResNet [16] as the backbone network, following established practices in recent facial expression recognition literature to ensure fair comparison with state-of-the-art methods that utilize similar pre-training strategies. The network was initialized with weights pre-trained on the MS-Celeb-1M dataset [17], a large-scale face recognition dataset containing over one million face images. This pre-training provides robust facial feature representations that transfer effectively to downstream expression recognition tasks. The training procedure was conducted over 100 epochs on the RAF-DB dataset with a batch size of 128 samples. We adopted the Adam optimizer with an initial learning rate of $1e-4$, chosen to provide stable convergence while avoiding overshooting during the optimization process. The learning rate was scheduled using ExponentialLR with a decay factor that gradually reduces the learning rate throughout training.

For the loss function weighting, both parameters λ and β were set to 1.0, indicating that the three loss components—classification loss, consistency loss, and attention alignment loss—contribute equally to the overall objective function.

V. RESULTS AND DISCUSSIONS

At a first glance, Figure 5 depicts the training and validation performance of the proposed model across 100 epochs. The validation curves closely follow the trends observed in the training curves. Leveraging pre-trained weights, the model converges rapidly, with training loss nearing zero by epoch 30. Validation curves stabilize after 20 epochs, but accuracy remains competitive (around 88%), demonstrating effective generalization to unseen data.

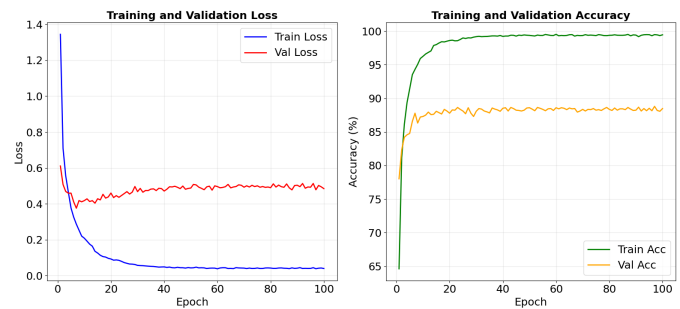


Fig. 5: Training Result

Comparison with State-of-the-Art. We benchmark MV-FER against state-of-the-art methods on the RAF-DB dataset using the evaluation setup described above. MV-FER is assessed in three variants, demonstrating progressive improvements through the integration of augmented views and AU-guided heatmap views. The quantitative results are reported in Table III. Our full three-view MV-FER framework achieves an accuracy of 89.54%, exceeding the recent strong published works: Guided Interpretable FER [5] (88.95%) by +0.59 percentage points and EAC [13] (88.75%) by +0.79 percentage points. Because these comparisons use the same ResNet-50 backbone as several competitors, the gains can be attributed to our multi-view learning design rather than architectural capacity.

Notably, the two-view variant (without any landmark information) already attains 88.75%, matching the best landmark-free baseline (EAC, 88.75%). When landmarks are introduced only as structured supervision (not as input features), MV-FER rises to 89.54%, outperforming all landmark-based methods, including pACNN [11] (85.07%), RAN [12] (86.90%), OADN [18] (87.16%), FERGCN [7] (88.23%), and Guided Interpretable FER [5] (88.95%).

Ablation within MV-FER: To demonstrate the effectiveness of each component in our framework, we conducted ablation experiments comparing three variants of our approach: (1) single-view baseline using only original images, (2) two-view variant incorporating original and augmented images, and (3) the full MV-FER framework including AU-guided heatmaps. The results reveal substantial progressive improvements: the single-view baseline achieves 85.43% accuracy, while the addition of the augmented view (horizontally flipped images) dramatically improves performance to 88.75% (+3.32 percentage points). The incorporation of AU-guided heatmaps in the full MV-FER framework further enhances accuracy to 89.54% (+0.79 percentage points), representing a cumulative improvement of +4.11 percentage points over the single-view baseline.

Confusion Matrix Analysis. An examination of the class-wise expression accuracy in Figure 6 reveals that the proposed model achieves strong recognition performance for several categories, most notably happiness (95.27%), neutral (90.44%), and sadness (88.08%). In contrast, fear (60.81%) remains the most challenging category, largely due to its similarity with expressions such as surprise and sadness. This outcome aligns

TABLE III: Comparison of methods on the RAF-DB dataset using overall accuracy (%).

No.	Method	Backbone	Landmark Used?	Accuracy
1	pACNN (IEEE Trans. Image Process.'19) [11]	ResNet50	Yes	85.07
2	LDL-ALSG (CVPR'20) [19]	ResNet50	No	85.53
3	RAN (IEEE Trans. Image Process.'20) [12]	ResNet18	Yes	86.90
4	SCN (CVPR'20) [20]	ResNet18	No	87.03
5	OADN (IJCB'20) [18]	ResNet50	Yes	87.16
6	DACL (WACV'21) [21]	ResNet18	No	87.8
7	FERGCN (MVA'22) [7]	ResNet18 + GCN	Yes	88.23
8	EAC (ECVA'22) [13]	ResNet50	No	88.75
9	Guided Interpretable FER (FG'24) [5]	ResNet50	Yes	88.95
10	MV-FER			
	(1) Single-view with only original images	ResNet50	No	85.43
	(2) Two-view includes augmented images	ResNet50	No	88.75 (+3.32)
	(3) Three-view includes AU-guided heatmaps	ResNet50	Yes	89.54 (+4.11)

*MV-FER is presented in three variants, showing progressive improvements when adding augmented and AU-guided heatmap views.

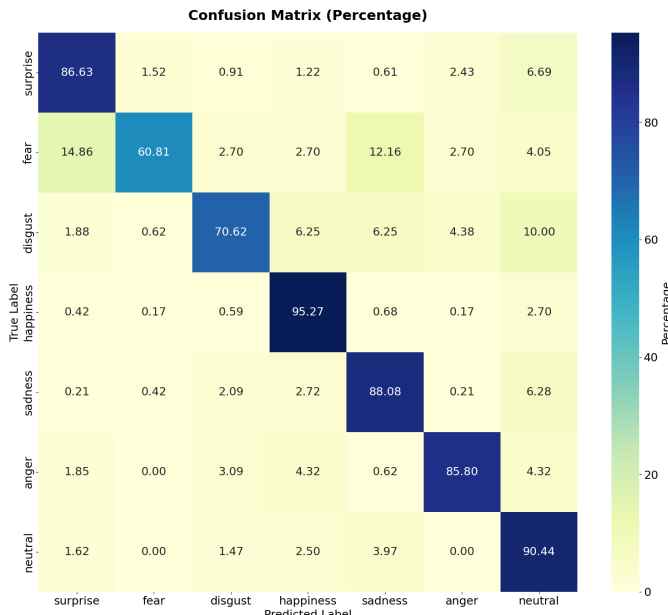


Fig. 6: Confusion Matrix Analysis (%) on RAF-DB

with prior findings in the FER literature, which highlight the inherent difficulty of distinguishing subtle expressions like fear and disgust from neighboring affective states [2].

VI. CONCLUSION

In this work, we introduced MV-FER, a novel multi view facial expression recognition framework that fundamentally differs from existing approaches in its systematic integration of domain-specific knowledge with multi-view learning paradigms. Instead of consuming AU heatmaps as inputs, we recast them as structured supervisory signals that align network attention with expression-discriminative regions, while a dual-branch architecture processes each image and its horizontal flip to enforce flip-semantic consistency.

Empirically, MV-FER achieves 89.54% accuracy on RAF-DB, surpassing recent strong published works that employ the same ResNet backbone, thereby demonstrating that performance gains arise from our multi-view modeling approach rather than architectural advantages. Ablation studies reveal that the incorporation of flipped views contributes the majority of the performance improvement, while AU-guided supervision provides consistent additional enhancement. While our approach currently relies on reliable landmarks for supervision, it opens clear paths for future work, including the development of self-supervised cross-view alignment mechanisms that could eliminate landmark dependencies while preserving the benefits of structured anatomical guidance. Future work may also explore emotion injection—generating or enhancing emotional expressions in facial images—to enrich data diversity, and extend the framework toward multimodal emotion analysis by extracting emotions from conversational audio or text for a more holistic understanding of human affect.

REFERENCES

- [1] B. Li and D. Lima, “Facial expression recognition via resnet-50,” *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, 2021.
- [2] H.-H. Dinh, H.-Q. Do, T.-T. Doan, C. Le, N. X. Bach, T. M. Phuong, and V.-V. Vu, “Fgw-fer: Lightweight facial expression recognition with attention,” *KSI Transactions on Internet and Information Systems*, vol. 17, no. 9, pp. 2505–2528, 2023.
- [3] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
- [4] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, “Automatic analysis of facial actions: A survey,” *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–347, 2019.
- [5] S. Belharbi, M. Pedersoli, A. L. Koerich, S. Bacon, and E. Granger, “Guided interpretable facial expression recognition via spatial action unit cues,” in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–10.
- [6] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2983–2991.

- [7] H. Meng, F. Yuan, Y. Tian, and T. Yan, "Facial expression recognition based on landmark-guided graph convolutional neural network," *Journal of Electronic Imaging*, vol. 31, no. 2, p. 023025, 2022. [Online]. Available: <https://doi.org/10.1117/1.JEI.31.2.023025>
- [8] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 3138–3147.
- [9] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "Poster++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, vol. 157, p. 110951, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320324007027>
- [10] X. Ning, X. Wang, S. Xu, W. Cai, Z. Liping, L. Yu, and W. Li, "A review of research on co-training," *Concurrency and Computation: Practice and Experience*, vol. 35, 03 2021.
- [11] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.
- [12] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.
- [13] Y. Zhang, C. Wang, X. Ling, e. S. Deng, Weihong, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 418–434.
- [14] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:6789015>
- [15] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [17] Y. Guo, L. Zhang, Y. Hu, X. He, e. B. Gao, Jianfeng, J. Matas, N. Sebe, and M. Welling, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 87–102.
- [18] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–9.
- [19] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 981–13 990.
- [20] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6896–6905.
- [21] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2401–2410.



Hoang Van Thanh is a researcher at Quang Binh University. He received a Ph.D. degree in Computer Engineering from University of Ulsan, Republic of Korea, in 2020. His research interests include pattern recognition, machine learning, natural language processing, and computer vision.



Hong-Quan Do is a lecturer at FPT University, Hanoi, and a PhD candidate at the Posts and Telecommunications Institute of Technology. His research focuses on clustering, computer vision, face analysis, and recommender systems. He is also actively involved in various e-Government projects and e-Commerce recommendation applications across both the public and private sectors.



Tu Minh Phuong is professor of computer science and Chairman of University Council at Posts and Telecommunications Institute of Technology. His research interests include machine learning, recommender systems, natural language processing, and computer vision. He received Ph.D. degree in control in technical systems from the National Academy of Sciences, Uzbekistan, in 1995.